哈尔滨工业大学计算机科学与技术学院

实验报告

课程名称：机器学习
课程类型：必修
实验题目：PCA模型
学号：1171000820
姓名：陈嵩

一、实验目的

实现一个PCA模型，能够对给定数据进行降维(即找到其中的主成分)

二、实验要求及实验环境

实验要求

测试

1. 首先人工生成一些数据（如三维数据），让它们主要分布在低维空间中，如首先让某个维度的方差远小于其它维度，然后对这些数据旋转。生成这些数据后，用你的PCA方法进行主成分提取。
2. 找一个人脸数据（小点样本量），用你实现PCA方法对该数据降维，找出一些主成分，然后用这些主成分对每一副人脸图像进行重建，比较一些它们与原图像有多大差别（用信噪比衡量）。

实验环境

- python 3.7.0

三、设计思想(本程序中用到的主要算法及数据结构)

1.算法原理

PCA(主成分分析，Principal Component Analysis)是最常用的一种降维方法。PCA的主要思想是将D维特征通过一组投影向量映射到K维上，这K维是全新的正交特征，称之为主成分，采用主成分作为数据的代表，有效地降低了数据维度，且保留了最多的信息。下面将叙述该算法的流程，以及背后对应的数学推导。

1.1 中心化

在PCA开始时需要对数据进行了中心化，即：对于数据集$\mathbf{D}=\{x_1,x_2,\dots,x_n\}$，其中$x_i$为$D$维变量。设样本的总数量为$N$。则$\mathbf{D}$的大小为$D*N$。

对于该数据集，中心向量或均值向量为

$$\mu=\frac{1}{N}\sum_{1}^{N}x_i \tag{1}$$

对数据集每个样本均进行如下操作：

$$x_i=x_i-\mu \tag{2}$$

之所以进行中心化，是因为经过中心化之后的常规的线性变换就是绕原点的旋转变化，也就是坐标变换；在后续的推导中，也会证明这一做法的必要性。

设使用的K个投影向量组成的矩阵为：

$$\mathbf{U}=\{u_1,u_2,\dots,u_k,\quad k<D\} \tag{3}$$

该矩阵的大小为$D*k$，且这些向量组成新的K维空间内的一组标准正交基。也就是任意i,j有

$$\begin{aligned} u_i^T u_i &= 1 \\ u_i u_j &= 0 \end{aligned} \tag{4}$$

1.2 最大方差推导

为简化推导，不妨设降维后的维度满足$K=1$。

我们的目的是最大化映射后的数据方差，对于样本x_i和均值μ，映射后的向量为：

$$\tilde{\mu} = u_1^T \mu \tag{5}$$

$$\tilde{x_i} = u_i^T x_i \tag{6}$$

新的方差为

$$\frac{1}{N}\sum_{n=1}^{N}\left\{u_1^T x_i - u_1^T \overline{\mu}\right\}^2 = u_1^T \mathbf{S} u_1 \tag{7}$$

其中S为协方差矩阵，有：

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)^T \tag{8}$$

因为需要满足$u_1^T u_1 = 1$，使用拉格朗日乘数法最大化方差，有：

$$u_1^T \mathbf{S} u_1 + \lambda(1 - u_1^T u_1) \tag{9}$$

对u_1求导，有：

$$\mathbf{S} u_1 = \lambda u_1 \tag{10}$$

因此，λ是$\mathbf{S}$矩阵的特征值，u_1是$\mathbf{S}$矩阵的特征向量时，满足取得极值条件。

对于$K>1$的情况，只需要选择其余的特征向量，即可满足对各个u的限制且取得极值。

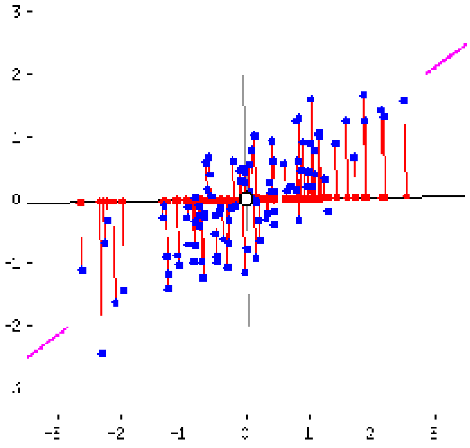
为了让方差尽可能的大，我们希望：

$$u_1^T \mathbf{S} u_1 = \lambda u^T u_1 = \lambda \tag{11}$$

尽可能的大，因此需要选择最大的前K个特征值所对应的特征向量即可。

1.3 最小误差推导

我们希望映射后的向量与原向量间的距离尽可能的小，对于降维后的数据，我们为其不足



这张图中，蓝色点是映射前的样本点，直线上的红色点是映射后的样本点，红色点所处的同一条直线随着所选择的投影方向不同而变化，在这个过程中，我们可以看出，蓝色点到映射后的样本均值中心白色点的距离是不变的，而在满足最大方差的前提下，红色点到中心点距离最大，根据勾股定理，此时满足蓝色点到红色点的距离（即红色线）的长度最小，满足映射前后的误差最小。因此两项要求达到的最优状态是一致的，也就是求解方式是等价的。

2.算法的实现

给定数据集$\mathbf{D}=\{x_1,x_2,\ldots,x_n\}$设样本矩阵为$\mathbf{X}$大小为$D*N$和想要降到的维数$\mathbf{K}$

1. 对数据集中的样本完成中心化：
   1. 计算样本均值$\mu=\frac{1}{N}\sum\limits_{n=1}^{N}x_n$
   2. 所有样本减去均值$x_n=x_n-\mu$
2. 计算协方差矩阵$\mathbf{X}^T\mathbf{X}$
3. 求出协方差矩阵$\mathbf{X}^T\mathbf{X}$的特征值
4. 取最大的$\mathbf{K}$个特征值对应的单位特征向量$v_1,v_2,\ldots,v_d$，构造投影矩阵$\mathbf{V}=(v_1,v_2,\ldots,v_k)$（这里使用字母$\mathbf{V}$与之前推导中不同）
5. 返回降维后的矩阵$\mathbf{XV}$此时的数据大小为$\mathbf{D}*\mathbf{K}$

此外，也可以使用奇异值分解来完成PCA。

对于样本矩阵$\mathbf{X}$进行奇异值分解，有：

$$\mathbf{X}=\mathbf{USV}^\top \tag{1}$$

而对$\frac{\mathbf{X}^T\mathbf{X}}{n-1}$进行特征值分解，有

$$\frac{\mathbf{X}^T\mathbf{X}}{n-1}=\mathbf{VLV}^\top \tag{2}$$

其中$\mathbf{L}$为对角阵，对角线上的值为特征值，$\mathbf{V}$即为与特征值相对应的各个特征向量组成的矩阵。

因此将(1)带入$\frac{\mathbf{X}^T\mathbf{X}}{n-1}$有

$$\frac{\mathbf{X}^T\mathbf{X}}{n-1}=\frac{\mathbf{VSU}^\top\mathbf{USV}^\top}{(n-1)}=\mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^\top \tag{3}$$

可以看出奇异值分解的结果中的U即为所求的$\mathbf{V}$,因此

$$\mathbf{XV}=\mathbf{USV^TV}=\mathbf{US} \tag{4}$$

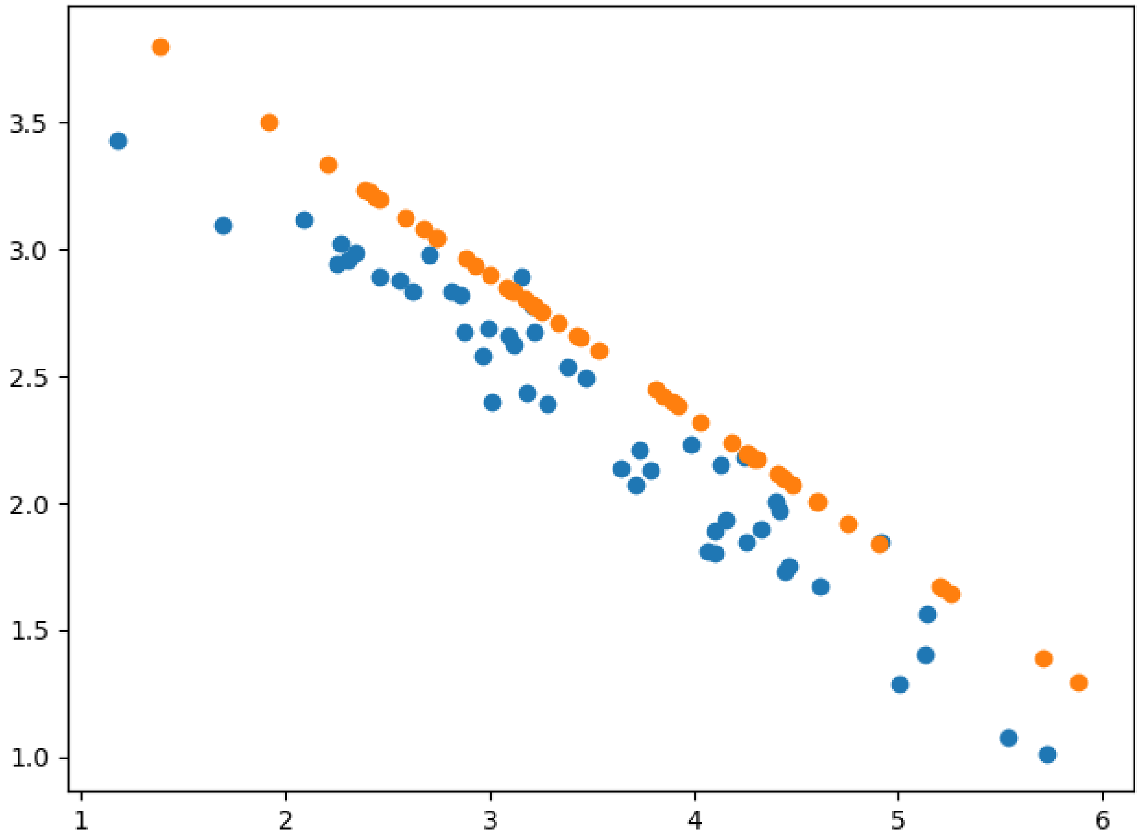
四、实验结果分析

1.生成数据的测试

为了方便进行数据可视化，在这里只进行了2维数据和3维数据的在PCA前后的对比实验。

2维数据的测试

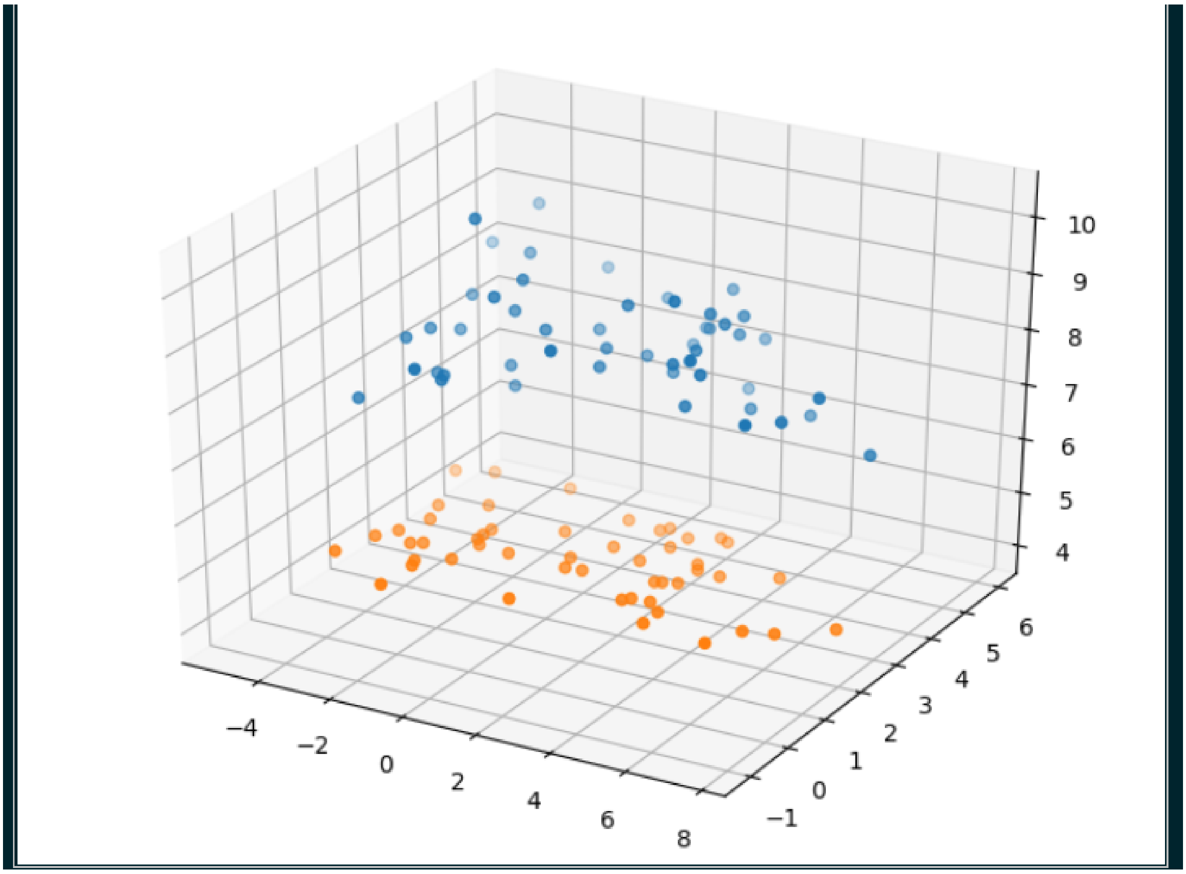
在2维数据的测试中，选择使用2维高斯分布产生样本，均值相同，第1维的方差远小于第2维的方差（$[1, 10] \ll [0.01, 0.10]$），对数据进行PCA降至一维，有：



可以看到在纵轴上数据散布较小（$[1.0, 3.5]$），数据分布主要依据于横轴上的值。

3维数据的测试

在3维数据的测试中，使用3维高斯分布随机产生样本，同样，第3维的方差远小于其余两个维度，PCA降维至两维有：



数据被投影到一个平面，数据主要依赖于x轴和y轴的分布，z轴被“舍弃”。

2.人脸数据集测试

[无背景带表情人脸数据]

数据样例：

jpeg格式，250*250像素，24bitcolor三通道存储。

PCA降维处理时转为灰度图像进行处理。一张图片相当于一个$250*250$的数据矩阵。

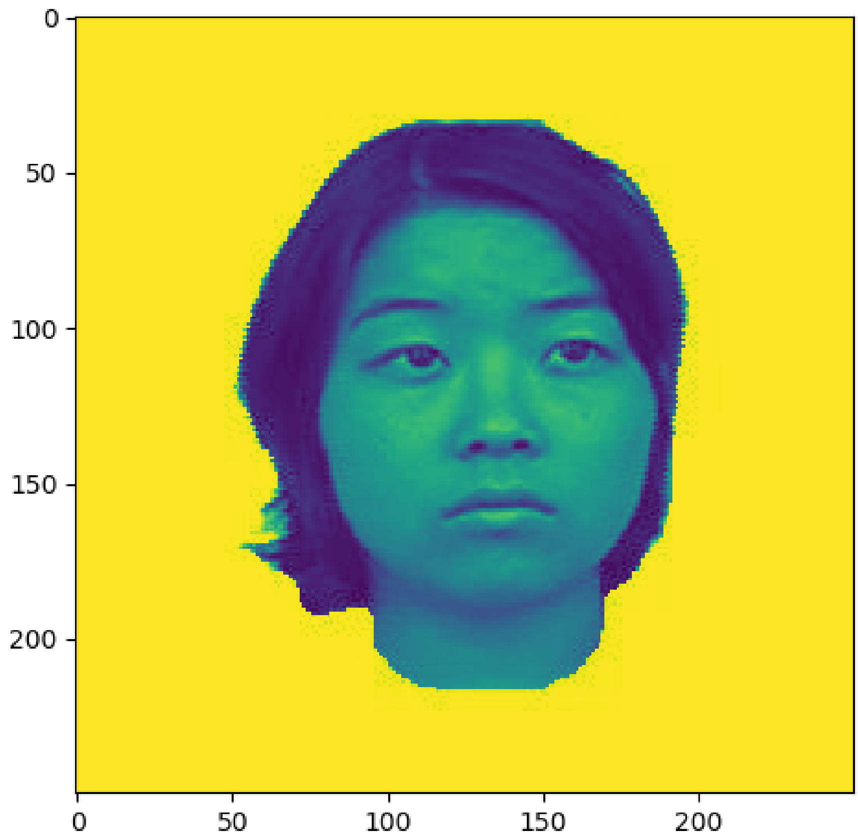
计算重建图片与原图片信噪比的公式为：

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}||I(i,j)-K(i,j)||^2$$

$$PSNR = 10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20\cdot\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

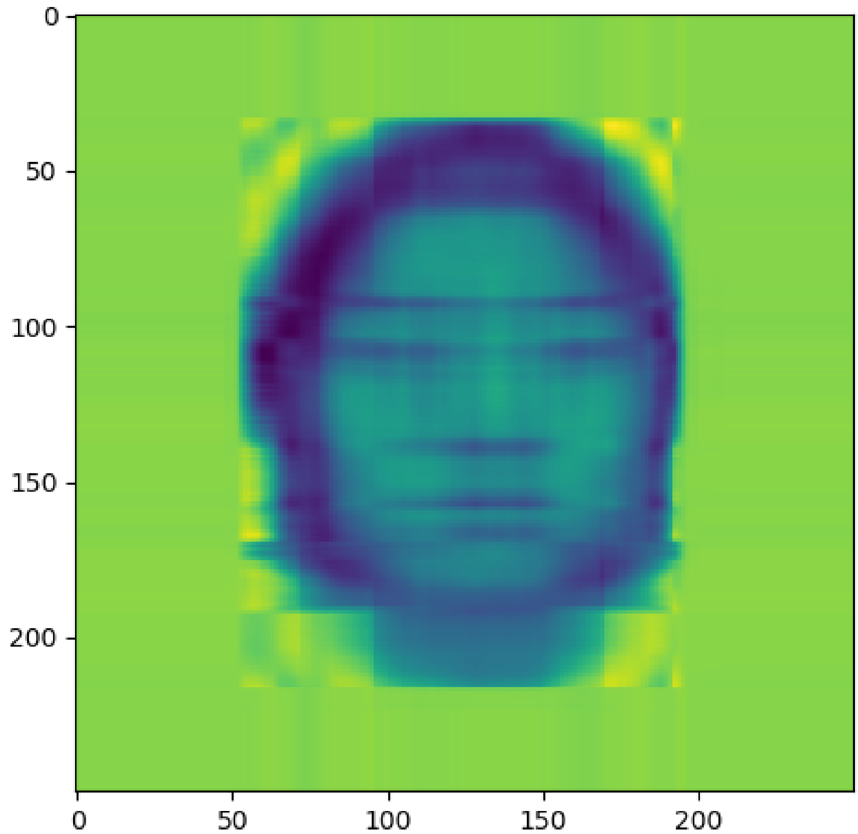
在保留维度$\mathbf{K}$选择不同值的情况下，有：

原图



$\mathbf{K = 5}$:

PCA降维后重建：

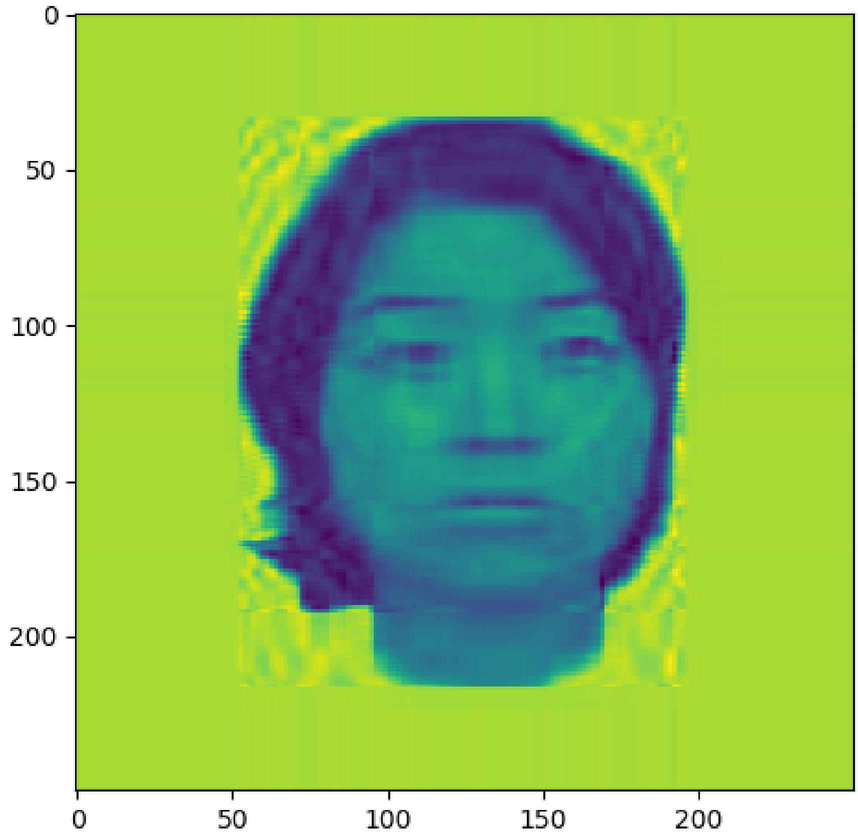


信噪比：21.69%

此时保留信息较少，图片较不清晰。

$\mathbf{K = 15}$

PCA降维后重建：

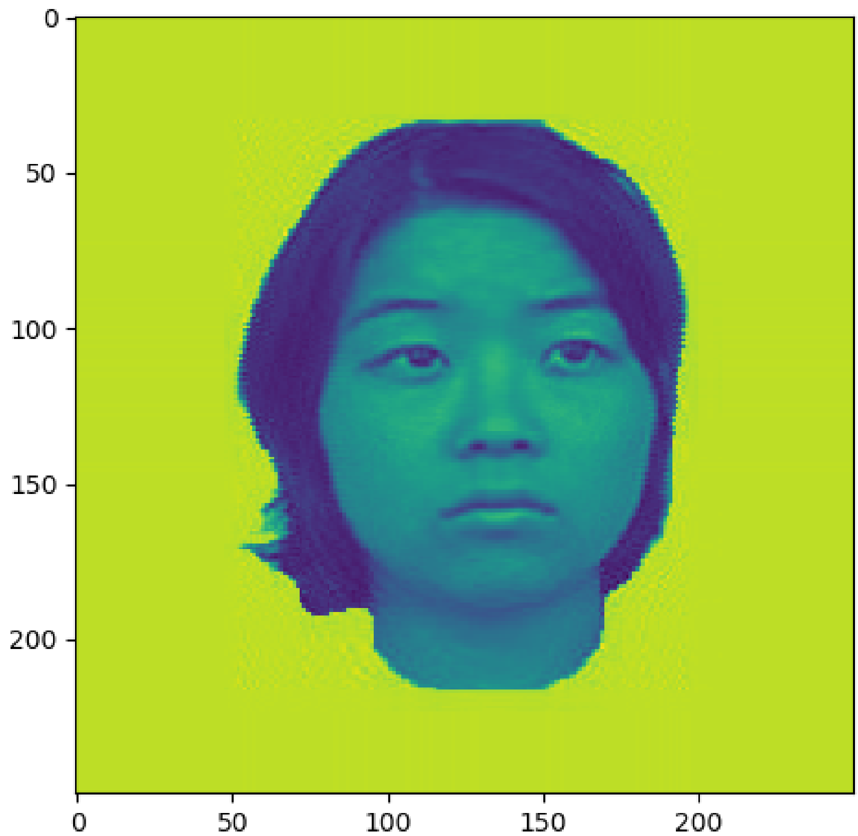


信噪比：27.70%

此时保留信息有所增加，图片清晰度增加。

$\mathbf{K = 50}$

PCA降维后重建：

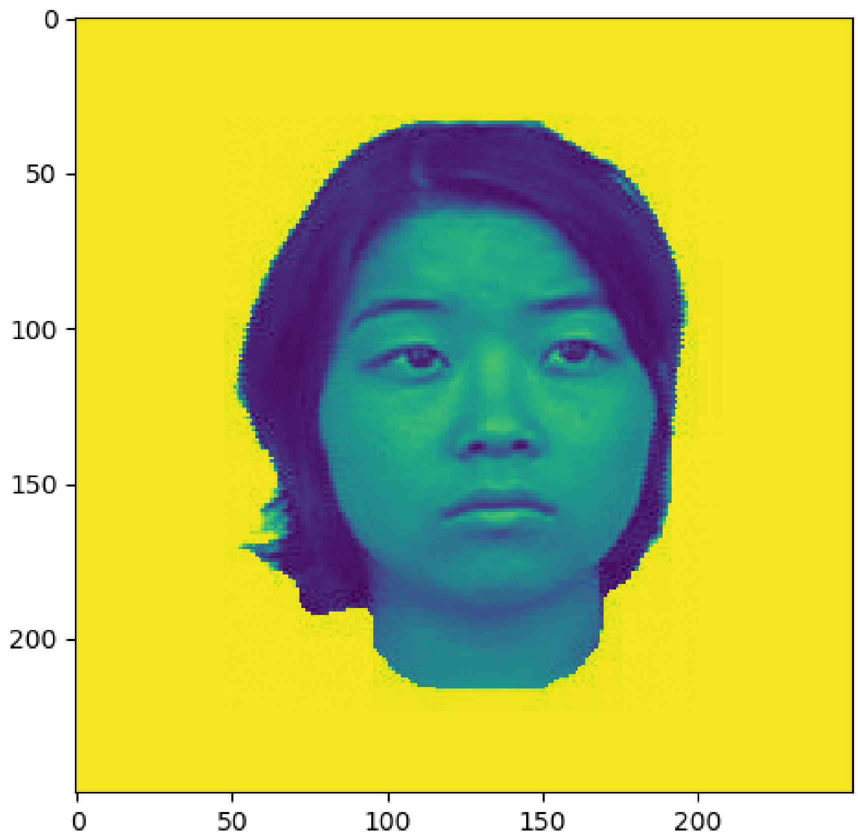


信噪比：37.21%

效果较好，人眼可察觉失真，但是可以接受。

$\mathbf{K} = 100$

PCA降维后重建：



信噪比：51.92%

效果最好，但图像质量提升有限，人眼仍能察觉到部分失真，说明增加的主成分对最终效果的贡献较小。

五、结论

- PCA算法选择协方差矩阵对应特征值最大的$\mathbf{K}$个特征向量作为基向量，将原数据在这几个方向上投影，将原本$\mathbf{D}$维的数据降至$\mathbf{K}$维，虽然会损失部分信息，但主要的信息存储在最大的部分里，投影后的数据储存了主要信息，且重建后主要依赖于较大的特征向量方向的值，达到了PCA推导时所采用的尽量保证最大方差和重建后的最小误差。有效降低了处理数据的复杂度，另外，舍弃特征值较小的特征向量，有利于对数据进行降噪，舍弃了部分噪声。
- PCA算法在人脸上表现较好，利用信噪比指标来衡量重建效果，可以在有效降低规模的同时保留较好的图像质量。

六、参考文献

- Face Place dataset
- Christopher Bishop. Pattern Recognition and Machine Learning.
- 周志华 著. 机器学习, 北京: 清华大学出版社, 2016.1

七、附录:源代码(带注释)

见邮件附件。